INTELLIGENT SYSTEMS IN ACCOUNTING, FINANCE AND MANAGEMENT
Intell. Sys. Acc. Fin. Mgmt. **22,** 227–247 (2015)
Published online in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/isaf.1376

TWITTER MINING FOR DISCOVERY, PREDICTION AND CAUSALITY: APPLICATIONS AND METHODOLOGIES

DANIEL E. O'LEARY*
Leventhal School of Accounting, University of Southern California, Los Angeles, CA, USA

SUMMARY

Twitter has found substantial use in a number of settings. For example, Twitter played a major role in the 'Arab Spring' and has been adopted by a large number of the Fortune 100. All of these and other events have led to a large database of Twitter tweets that has attracted the attention of a number of companies and researchers through what has become known as 'Twitter mining' (also known as 'TwitterMining'). This paper analyses some of the approaches used to gather information and knowledge from Twitter for Twitter mining. In addition, this paper reviews a number of the applications that employ Twitter Mining, investigating Twitter information for prediction, discovery and as an informational basis of causation. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: analytics; big data; business intelligence; efficient markets; event studies; social media; Twitter mining

1. INTRODUCTION

Twitter has become a critical social media tool that has a number of important capabilities, such as communication, building community and collective action organization. Reportedly, there have been over 300 billion tweets sent as of 13 October 2013.[1] Unlike newswire sources, Twitter tweets go beyond factual information to provide a wide range of public opinion on a topic. Tweets also contain jokes, rumour, commentary and opinion. Tweets often take information that is only distributed in some local area and expand that diffusion to broader areas.

As a result of these capabilities, Twitter is helping to change society and business, as illustrated by the following developments.

- As of 13 August 2013, there were a reported 500 million tweets per day,[2] up from June 2011, it was reported that there were over 200 million tweets per day and [3]up from 2 million per day in January 2009. As a result, Twitter is huge and continues to rapidly grow over time.
- Twitter and Facebook were key tools for communicating and generating collective action in the recent 'Arab Spring' as they were used to 'organize protests' or 'spread awareness of protests' (e.g., Huang, 2011).
- During the massive earthquake in Japan, when landlines and mobile phone lines got stuck much communication was done using Twitter, including by emergency services (Nguyen *et al*., 2011).
- Apparently, Twitter is broadly used in the USA by business. For example, Swartz (2009) noted that more than half of the Fortune 100 use Twitter for a range of activities, including customer service,

* Correspondence to: Daniel E. O'Leary, 3660 Trousdale Parkway, University of Southern California, Los Angeles, CA 90089-0441, USA. E-mail: oleary@usc.edu

[1] http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/#.UtVsf9GA1aQ.
[2] https://blog.twitter.com/2013/new-tweets-per-second-record-and-how.
[3] http://blog.twitter.com/2011/06/200-million-tweets-per-day.html.

recruiting and news dispersion. Wexler (2013) suggested that the entire Fortune 500 will be Twitter users.
- Twitter is the original source of many news releases. Further, in some cases it has been used to send information that conventional news media was wrong. For example, although the death of American football coach Joe Paterno had been announced by major news services, his son used Twitter to indicate that the news '… report is wrong'. As a result, Tweeter tweets serve to dis-intermediate original news sources and consumers of that information.
- Major news sources, such as CNN, have analysed the impact of Twitter (e.g. Sutter, 2009) because of its popularity, suggesting Twitter has an important impact on society.

Because of Twitter's importance, Lohr (2010) noted that the Library of Congress in the USA has decided to archive Twitter microblogs. Apparently, this effort is part of the 'Web Capture' project at the library begun to capture information about significant events.

1.1. Twitter Mining

Since Twitter has become so important, one perspective is to treat Twitter exchanges as 'data' and ultimately mine that data for its potential content. Accordingly, researchers have begun to apply classic text mining approaches and they have begun to leverage some of the unique features of Twitter to gather the increasing amounts of knowledge out of the tweets. Further, researchers have begun to analyze how different events manifest themselves in Twitter tweets. behavior.

There has been increasing attention focused on so-called 'Twitter-mining'. As an example, in January 2012, I found zero results under Google Scholar for the term 'Twittermining' and only 22 occurrences of the term 'Twitter Mining'. By July 2012, those numbers had increased to 1 and 28 respectively. In January 2014, there was one under 'Twittermining' and 87 under 'Twitter Mining'. In July 2015, there were two entries under 'Twittermining' and 197 under 'Twitter Mining'. Accordingly, the purpose of this paper is to review and extend the literature associated with the notion of Twitter mining/Twittermining focusing on methodologies and applications and the role of Twitter information in prediction, discovery and causation.

This paper is consistent with a business intelligence or knowledge discovery view of the firm, where intelligence is sought from virtually all signals available to an enterprise. In particular, Twitter is seen as providing a source of information to support decision-making whether that decision-making is about marketing decisions, political decisions or other concerns. In addition, this paper is consistent with Twitter as so-called 'big data' (e.g., O'Leary, 2013a,b). Twitter provides high-volume, high-velocity and high-variety unstructured data that can be used to support decision-making.

1.2. Purposes of this Paper

This paper has a number of different purposes. First, this paper brings together a number of different references related to capturing information from Twitter. Second, this paper provides an investigation of the different types of information that can be captured from Twitter tweets, and so on. Third, this paper provides a summary of a number of different types of applications made using Twitter data, including predictions, discovery and analysis of potential causation. Fourth, this paper provides an overview of some methodological issues involved in the analysis of Twitter information, including, for example, a review of some potential lexicons for sentiment analysis.

1.3. This Paper

This paper proceeds in the following manner. Section 1 provides the motivation and a background for the paper. Section 2 provides a brief background into Twitter and Twitter mining, investigating some of the unique issues associated with Twitter and Twitter messages. Section 3 briefly analyses some artificial intelligence approaches that could be used to analyse Twitter message text. Sections 4–6 investigate some of the characteristics of Twitter that can be mined. Section 7 analyses the findings of some of the previous studies aimed at using Twitter data to discover potential knowledge. Section 8 investigates the use of Twitter to predict, while Section 9 investigates the use of Twitter from the perspective of causation. Finally, Section 10 summarizes the paper, examines the contributions of the paper and analyses some potential extensions.

2. TWITTER BACKGROUND

Reportedly, Twitter was launched on 21 March 2006.[4] Currently, there are over 300 million active Twitter users.[5] Twitter is a microblogging tool where users can send messages of up to 140 characters. Individual messages are referred to as ‘tweets’. People can ‘follow’ others or people can be ‘followed’. In those cases, when there is a tweet, the message is directed to all who follow a particular participant.

Since Twitter messages are relatively short, the messages may use abbreviations, and the grammar is not likely to be correct. These issues can limit classic translation or news processing approaches that depend on correct spelling, syntax and sentence structure. In addition, messages are likely to include other characters not typically used in normal communications; for example, ‘#’ for hashtags and ‘@’ as part of the user name.

2.1. How is Twitter Used?

Initially, Twitter asked its users, ‘What are you doing?’ However, in 2009, they changed that official question to ‘What’s happening?’[6] Thus, at the most basic level Twitter is a social media, a microblogging tool to answer those questions. As a result, it probably is not surprising that Java *et al*. (2007) suggest that using Twitter people talk about their daily routine, have conversations and share information.

Twitter and microblogs differ from other blogs in at least two ways, as noted by Java *et al*. (2007):

> Compared to regular blogging, microblogging fulfills a need for an even faster mode of communication. By encouraging shorter posts, it lowers users’ requirement of time and thought investment for content generation. … The second important difference is the frequency of update. On average, a prolific blogger may update her blog once every few days; on the other hand a microblogger may post several updates in a single day.

With Twitter, people often know the people they are sending the message to. If people are friends, then Twitter provides another vehicle of communication and affirmation of that friendship.

However, Twitter has provided a critical news function in a number of settings. As a result, some researchers have analysed the issue as to whether Twitter is a news media or a social media (e.g., Kwak

[4] http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/#.UtVvF9GA1aQ.
[5] http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.
[6] http://mashable.com/2009/11/19/twitter-whats-happening/.

DOI: 10.1002/isaf

et al., 2010)? Twitter messages announce and discuss a range of discussions that are often aimed at generating or passing on news stories about current events. Twitter is often the source of news in sports. For example, many professional and college athletes have Twitter accounts that are 'followed' by a number of others. In addition, Twitter also has been an important news source for other events, such as the American student jailed in Cairo,[7] the Japanese tsunami and others.

Ediger *et al*. (2010) find that Twitter messages can be assembled as a tree capturing news dissemination. They also find that much of the activity in blogs and microblogs such as Twitter is the result of (redundant) rebroadcasting of previous information.

Although Twitter tweets are often bits of news, they can capture other information. Oftentimes they contain rumours, gossip, opinions, commentaries, and so on, generating substantial sentiment about a wide range of issues. As another example, Ediger *et al*. (2010) not only found news propagation, but also found clusters of conversations about more personal concerns.

2.2. What is 'Twitter Mining'?

Twitter mining is analysing Twitter message information to predict, discover or investigate potential causation. Twitter mining includes text mining designed to specifically leverage Twitter tweet content and contexts. Twitter mining can include analysing additional information associated with tweets, including names, hashtags and other characteristics. Twitter mining also employs the substantial quantitative information (numbers of tweets, retweets,, likes, favorites, etc.) to try to better understand the phenomena under consideration. Finally, Twitter mining can examine how Twitter tweets, retweets, etc., capture and reflect different events or even how Twitter relates to other social and conventional media.

As noted above, by late 2013, Twitter had over 500 million messages per day. As a result, it is impossible for a human to read and analyse anything but a small percentage of those messages. Thus, it is important to develop additional computer-based resources and approaches designed to facilitate examination and analysis of those messages. Accordingly, Twitter mining uses a range of different approaches, including data mining, to investigate the content of Twitter with the purpose of finding information or knowledge about various products, individuals, organizations and concepts.

In addition, Twitter mining often is concerned with providing structure to the unstructured information content in Twitter, such as capturing the sentiment of Twitter tweets. For example, Nguyen *et al*. (2011) suggest that Twitter is a 'sensor' of the real world, since so much daily and emergency activity is done using Twitter. From the sentiment analysis perspective, messages can be analysed for their positive or negative (or neutral) sensing of the world.

2.3. Twitter Use Differs by Country

There are roughly 65 million active Twitter users in the USA, and 239 million international users.[8] As a result, it is not surprising that a number of researchers have found that the volume and the nature of tweets seem to vary by country. As example, in a case study analysis that included tweets from France, Germany, Spain and the Netherlands, Dijikman *et al*. (2015) found the Netherlands had the highest and France the lowest per capita use.

[7] http://www.washingtonpost.com/blogs/blogpost/post/luke-gates-american-student-arrested-in-cairo-wrote-on-twitter-of-wanting-to-die-in-egypt/2011/11/22/gIQA61Y3kN_blog.html
[8] http://www.statista.com/statistics/274565/monthly-active-international-twitter-users/

DOI: 10.1002/isaf

2.4. Who Can Have a Twitter Account?

Almost anyone can have a Twitter account. Organizations or people affiliated with organizations can have accounts. As a result, there is an asymmetry of information in that people may or may not be affiliated with an organization, but that organization-affiliated information about them will not necessarily be publicly available. As a result, people who appear to tweet or retweet positively for an organization may actually be affiliated with that organization. People that tweet aggressively against some organization may be associated with competitors to those organizations. Accordingly, this can make analysis of Twitter data difficult. In particular, it is difficult to divide the set of Twitter users into those users affiliated with organization X and those not affiliated with organization X. As a result, it can be difficult to tease out differences between such groups of users. Thus, Twitter data may contain certain biases that can confound the results, based on the nature of the questions examined.

3. ARTIFICIAL-INTELLIGENCE-BASED APPROACHES TO SIMILAR SETTINGS

There is substantial previous research in related areas of artificial intelligence focused on reading and understanding stories and text, and identifying events in conventional media. This section provides a brief summary of some of that research. Since the focus of this paper is on Twitter mining, the scope of this review is limited to a few sources.

For example, a number of systems have been developed and designed to read and understand news stories from text:

- DeJong (1979) developed FRUMP (Fast Reading Understanding and Memory Program) that was designed to skim news stories and then produce a summary of what it understands.
- Liebowitz (1980) developed a system that learns about the world by reading stories from newspapers and news wires making generalizations and it uses those generalizations to help in understanding future stories.
- Hayes and Weinstein (1990) developed a system that would index the content of a database of news stories.
- Mueller (2002) surveyed much of the story-understanding literature, while Mueller (2004) investigated common sense reasoning as a basis to understand news stories involving terrorism. Mueller (2002) argued that classic story understanding does not scale well and is suitable primarily for smaller scale problems.

In addition, there have been a number of systems designed to recognize and identify events from text:

- Allan *et al.* (1998) and Papka (1999) investigate monitoring a stream of broadcast news in order to identify new events.
- Vargas-Vera and Celjuska (2004) focused on event recognition in news stories and used an ontology to extract knowledge from those news stories.
- Westermann and Jain (2007) examined development of a common event model for multimedia applications that could be used for a number of purposes, including search and mining of events.

Rather than replicate all of these capabilities in a Twitter environment, we acknowledge the existence of these and many other approaches and capabilities and instead focus on relatively unique aspects of Twitter.

DOI: 10.1002/isaf

4. WHAT IS AVAILABLE TO MINE?

Although Tweets are at most 140 characters, Twitter provides a range of potential information to mine. There is substantial quantitative information that includes statistics such as number of tweets, number of retweets, number of followers and number of people that are being followed and other statistics. In addition, there is substantial non-numeric qualitative information in terms of the actual messages. Using this data can provide insights; however, there are also unique challenges associated with Twitter messages.

4.1. Hands-On Approaches to Twitter Mining

Some sources have discussed issues associated with direct examination of Twitter as a basis for Twitter mining. Russell (2013) examines a number of hands-on approaches for extracting a range of information. As another example, Russell (2011) offers 21 'recipes' for mining Twitter. Each recipe has some code that provides the user the ability to do some analysis of Twitter data.

4.2. Number of Tweets

Perhaps the most straightforward data source in Twitter is the sheer number of tweets; for example, about a particular event, agent, resource or location. For example, Asur and Huberman (2010) found that the rate at which movie tweets are generated can be used to build a model for predicting movie box office revenues. Further, they found the resulting models were better than models based on prediction markets such as the 'Hollywood Stock Exchange' (HSX.com).

4.3. Dates of the Tweets

Each tweet is date stamped. As a result, date may be an important characteristic that can be mined. For example, in the analysis of a political issue, O'Leary (2012) found that users from different political camps executed their tweets and retweets at different times, reflecting different events of importance to those camps.

4.4. Number of Retweets

The number of retweets provides a measure of the interest in a particular tweet. A number of researchers have examined retweets that are seen as a key mechanism for distributing information in Twitter. Suh *et al*. (2010) find that the presence of different types of information in the message is related to the extent to which a tweet is retweeted. For example, messages with embedded links are more likely to be retweeted.

4.5. Number of 'Favorites'

If a message is of particular interest to someone who sees the message they can label that message a 'favorite'. As a result, the extent to which a message has been labelled a favourite provides a measure of the interest in the particular tweet.

DOI: 10.1002/isaf

4.6. Following and Followers

Twitter allows users to ‘follow’ other users. These follower relationships allow the building of networks of interacting users. As an example, Conover *et al*. (2012) use Twitter messages to develop political information diffusion networks and use those networks to predict the political alignment of Twitter users. In another example, of following–follower analyses, Huberman *et al*. (2008) found a ‘… study of social interactions within Twitter reveals that the driver of usage is a sparse and hidden network of connections underlying the “declared” set of friends and followers’.

4.7. Location Information

Twitter users can enable location services.[9] Users can toggle on the ‘Share precise location’ button and their precise latitude and longitude will be associated with the tweet. Otherwise, users can attach a location (such as a city) of their choice to the tweet. In addition, other researchers have begun to infer location information. For example, Jurgens (2013) found that social networks help infer location of the user, with roughly 50% of a network within 10 kilometers of the user. In addition, Jurgens (2013) finds that Twitter is useful in finding other social media applications of the particular user.

Location information is important in the analysis of different types of problems. For example, later in the paper I examine disease diffusion and food poisoning discovery that both can employ location information.

4.8. Hashtags

Hashtags, provided by the message sender, indicated using ‘#’, can be used for a number of reasons. As noted by Twitter,[10] the hashtag was used as a way of categorizing messages, capturing keywords or topics. Twitter also suggests that hashtags help those terms show more easily when using Twitter Search. Hashtags also can provide a feeling or sentiment to a message. Users typically place hashtags at the end of the tweet; however, they can occur anywhere in a tweet. Hashtags have become so integrated into contemporary culture that there are even comedy sketches about them (https://www.youtube.com/watch?v=57dzaMaouXA).

5. MESSAGE SEMANTIC CONTENT

Twitter messages contain different semantic content that can be used to facilitate gathering the meaning and other message characteristics.

5.1. Qualitative Information

Since tweets are text, one approach is to text mine content for particular concepts. Typically this has been done by investigating tweets individually. For example, Nishida *et al*. (2011) provide an approach for classifying arbitrary tweets as being an interesting topic. As another example, Zhang *et al*. (2011)

[9] https://support.twitter.com/articles/78525#.

[10] https://support.twitter.com/articles/49309-using-hashtags-on-twitter#.

DOI: 10.1002/isaf

investigated a guided search for specific information and an unsupervised search for hidden topics underlying the tweets.

Nguyen *et al*. (2011) parse each individual tweet separately and generate a semantic network based on Twitter messages, capturing intelligence from human activities as events occurred; for example, the Japan earthquake. In their analysis of the earthquake they used a model based on 'activities', where the key elements of an activity are actor, action and object. In addition, they argued that it was important to know where and when an activity occurred; thus, they added time and location to their model, ultimately generating a schema of information. Such issues are particularly important in extreme event emergency settings where those variables might facilitate life-saving activity.

Teufl and Kraxberger (2011) investigated Twitter data on the Egyptian revolution. They argued that in order to extract knowledge, there were three key requirements. First, there was a need to be able to extract knowledge in layers going from more aggregate to more detailed. Second, they indicated that once data was extracted it needed to be represented, and that representation could take numerous formats, ranging from lists to visualizations of maps of tweets. Third, they indicated there needs to be a convenient user interface to allow access to multiple layers and to the knowledge gathered in the analysis. In their analysis of the revolution they noted a number of different events, including bombing of a church, starting the protests, arrests and clashes with police, involvement of the army and others.

5.2. Sentiment

There is substantial research on analysing the sentiment associated with Twitter messages (Go *et al*., 2009; Thelwall *et al*., 2011; Yalamanchi, 2011). Typically, the analysis is based on key word analysis of tweets, where it is assumed that words have a positive, neutral or negative sentiment. That approach often employs dictionaries (lexicons) of designated words. For example, positive words are likely to include words like 'love', 'wonderful' or 'great', while negative words can include 'bad', 'stupid' and 'waste'.

Capturing sentiment in Twitter (and other social media) requires capturing a number of non-dictionary symbols and other issues (Agarwal *et al*., 2011). In particular, sentiment will need to be applied to emoticons, abbreviations, repeated characters (extended words), abbreviations and other symbols.

- Emoticons can provide indications of both positive and negative (☺ and ☹) emotions.
- Since Twitter has been limited to 140 characters or less, users have come up with a number of abbreviations. For example, 'gr8' (great), 'bff' best friend forever, lol (laughing out loud) and others.
- Punctuation can help provide meaning in Twitter. However, such punctuation may exceed or not conform to traditional use (e.g. '!!!').
- Oftentimes a word is extended to convey a particular sentiment; for example, 'cooool' for the word 'cool'. However, it is difficult for a dictionary to anticipate how any different 'o's will be in the extended word.
- Numbers can be representative of particular concepts. For example, each of the following phrases derives important information from the particular numbers: 'We are #1!' 'She is a 10!' 'What a 0!'

A summary of some sites that provide Twitter sentiment analysis is given in Table I.

Table I. Some links for Twitter sentiment analysis

Name	Link
Rankspeed	http://www.rankspeed.com/
Social Mention	http://socialmention.com/
Twitter Sentiment	http://www.sentiment140.com/

5.3. Sentiment Lexicons

There are a number of different lexicons available. Perhaps the most general is WordNet (Fellbaum, 1998). A lexicon aimed at sentiment for financial applications is given by Loughran and McDonald (2011). Still other lexicons aimed at capturing sentiment have been investigated by Mohammad *et al*. (2009), Hu and Liu (2004), Baccianella *et al*. (2010), Wilson *et al*. (2005) and Stone *et al*. (1966). A summary of some available lexicons is given in Table II.

5.4. Named Entities

Shen *et al*. (2013) investigated approaches to linking entities named in Twitter messages. For example, concern might be with identifying messages associated with 'Tony Allen', a player in the National Basketball Association. Their approach uses a knowledge base and information derived from the content of the specific tweet and by analysing the user's interests as specified in other tweets. Ultimately, they build a graph linking different entities generating weights on the links to capture the strength of the interdependence.

5.5. Context

However, analysing individual messages does not take into account critical context information. For example, it is likely that who issues the message, what the message is about and when it is issued are likely to set a context that could be used to eliminate potential ambiguities, and provide the context to facilitate sentiment of Twitter messages. Further, one message can be part of a larger dialogue that could be used to provide 'meta' sentiment.

In some Twitter messages the user embeds a link providing the ability to generate greater detail. As an example, 'CNN Breaking News' includes a link to a more detailed story version. Accordingly, if the Tweet is ambiguous, the attached link can provide context disambiguating information.

Table II. Some lexicons for sentiment analysis

Lexicon	Link
WordNet	https://wordnet.princeton.edu/
Loughran and McDonald Financial Sentiment Dictionaries	http://www3.nd.edu/~mcdonald/Word_Lists.html
MSOL	http://saifmohammad.com/Lexicons/MSOL-June15-09.txt.zip
Opinion Lexicon	http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon
Sentiwordnet	http://sentiwordnet.isti.cnr.it/downloadFile.php
Subjectivity Lexicon	http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
The General Inquirer	http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm

DOI: 10.1002/isaf

6. ADDITIONAL SOURCES OF TWITTER INFORMATION

Twitter also includes substantial 'header' information that can be useful in assessing a range of issues about the Twitter messages. If there is ambiguity associated with the content, then names, descriptors and embedded links can provide some ability to disambiguate and establish relevance and meaning of a message as part of the mining process.

6.1. Names

In some cases the Twitter names of the participants can help guide 'semantic understanding' of the Twitter tweets (e.g. setting expectations as to the content of the messages). For example, 'CNN Breaking News' establishes both a source and a set of expectations of the information in the tweets: 'CNN' is the source and 'Breaking News' provides the basis of the exchange.

In addition, some names provide insight into the basic sentiment likely to be found in the tweets, establishing a set of expectations. For example, the name 'Chucknicecomic' has three indicators of the content of the specific person. 'Chuck' suggests that the Tweeter is a male, 'nice' suggests that the person and their tweets will have a particular disposition or last name, and 'comic' suggests that the person's tweets will be humorous. Each of these factors provides 'expectations' that can be substantiated, or not, in a manner consistent with previous research in artificial intelligence analysis of text.

6.2. Name Descriptors

Oftentimes, in addition to the name of the person doing the tweets, there is a descriptor and there often is semantic information in that descriptor. For example, 'CNN Breaking News' notes that 'CNN.com is among the world's leaders in online news and information delivery'. Such descriptors provide context information that also may be useful in disambiguating meaning of tweet content by providing a setting.

7. DISCOVERY USING TWITTER MINING

Twitter has been used to 'discover' different phenomena (descriptive analytics that could signal certain events) and information about the nature of different phenomena that could support decision-making and provide business intelligence, including the following.

7.1. Accidental or Careless Activity

Twitter can be analysed for accidental disclosures. For example, messages could be monitored to determine sayings such as there will be a 'party after our earnings announcement' disclosing positive sentiment related to earning expectations, but an inappropriate disclosure for publicly held companies. As another example, informally, I was told of one company that posted a picture of a project team working late to develop a new strategy. Unfortunately, the resulting strategy was published on a blackboard behind the people in the picture, ultimately giving their competitor immediate deep insight into their new strategy.

7.2. Emergency/Disaster Situations

There has been substantial analysis of Twitter in emergency and disaster situations, such as hurricanes, earthquakes and other events. Hughes and Palen (2009) found that Twitter messages sent during emergency situation events are different from general Twitter use. They find that there is more information broadcasting and brokerage, while general Twitter use offers more of an information sharing purpose. Vieweg *et al*. (2010) found that Twitter was able to help enhance 'situation awareness' (the big picture) during two different emergency situations that occurred during 2009.

Mendoza *et al*. (2010) investigated the ability of Twitter participants to discriminate between false rumours and confirmed news in emergency situations, such as an earthquake. In addition, they investigated how users tried to filter the false news from accurate news, because retweets often were not reliable information. Similarly, Acar and Muraki (2011) found that after the Japanese tsunami retweets often were not reliable. As a result, they found that users tried to use official hashtags and limit the number of retweets.

Mandel *et al*. (2012) examined tweets that occurred around the time of hurricane Irene. They found that the number of messages peaked at around the time the hurricane hit particular regions. In addition, they found that concern varied based on region. These results suggest the importance of location information.

7.3. Sources of Food Poisoning (#foodpoisoning)

Twitter messages are now being monitored to discover the potential existence of food poisoning at restaurants. As noted by Maron (2014) in Chicago, Twitter tweets led to the closing of 21 restaurants and another 33 were forced to fix violations. In a closely related report, Fox (2014) noted that Yelp reviews were being used by New York City to find food poisoning at restaurants. For such discoveries, location information can be particularly helpful, but frequently sufficient descriptor information is available in the tweet to determine the specific restaurant.

7.4. Political Events

O'Leary (2012) investigated the US Senate use of Twitter in an analysis of the Protect Intellectual Property Act. As part of protesting that act, many sources on the Internet 'shut down'. For example, Wikipedia went off line for a day. In order to analyse the Twitter data, O'Leary partitioned the senators into three groups and analysed the Twitter behaviour of each group: nonsponsors, co-sponsors and former co-sponsors. O'Leary found a number of differences in the different groups' Twitter messages, resulting in different timing, quantities and number of retweets as the three groups responded differently to the same events.

7.5. Fraud

O'Leary (2011) investigated the notion of using data from sources like Twitter as a means of attempting to identify fraud and other concerns. In particular, O'Leary (2011) suggested examining Twitter and other social media sources for evidence of misuse of assets, information about frauds, and so on.

7.6. Identify Influential Contributors

Weng *et al*. (2010) examined the issue of determining the most influential users of Twitter. In so doing, they generated a version of 'PageRank' referred to as 'TwitterRank' to measure the influence of different users in Twitter. In their analysis they found substantial 'reciprocity'. For example, they noted that 72.4% of the users follow more than 80% of their followers, and 80.5% of the users have 80% of users they are following follow them back. Cha *et al*. (2010) expanded the notion of influence beyond followers to consider other variables, including retweets and 'mentions' (number of times a user's names is included in a message). Bakshy *et al*. (2011) found that determining the most influential Twitter users is 'unreliable'. As a result, they recommend that if it is important to generate influence, then it is important to target '… large numbers of potential influencers'.

7.7. Reputation Management

A number of researchers have explored using Twitter as part of 'Online Reputation Management'. In these systems, typically, Twitter messages are scanned to determine if the message refers to some particular entity, such as a company or individual. For example, Jansen *et al*. (2009) analysed more than 150,000 Twitter tweets and found that roughly 20% contained information about a particular company or product. Of that 20% roughly 50% were positive messages, while 33% were critical of the company or product. Accordingly, it can be critical for companies to continuously monitor information from Twitter to understand the nature of their customers' concerns and preferences.

An alternative approach is to try to manage the reputation. Prokofieva (2014) finds that companies can use Twitter to attract an investor's attention and decrease information asymmetries. She also finds that there is an abnormal difference between the bid–ask spread and the number of tweets issued by the company during the earnings announcement period.

8. PREDICTION USING TWITTER MINING

Twitter has been used frequently as a data source for predicting certain sets of events (i.e. predictive analytics). In some cases Twitter tweet variables are the only variable used in the analysis, while Twitter variables are increasingly only one of the types of variables of interest. A wide range of prediction models based on Twitter variables have been made in a number of areas, including the following.

8.1. Predicting Elections

There has been substantial research worldwide analysing whether variables such as the frequency of occurrence of the names of different candidates and the names of political parties can be used to make predictions about votes in elections.

DiGrazia *et al*. (2013) find that reliable data about political behaviour can be captured from social media. In particular, they find a statistically significant association between tweets that mention a candidate for the US House of Representatives and their subsequent electoral performance. Conover *et al*. (2012) use Twitter messages to develop political information diffusion networks, and use those networks to predict the political alignment of Twitter users. Networks based on followers and following can be constructed to facilitate understanding of diffusion of information through the networks. In addition, Conover *et al*. (2012) used hashtags as a basis of predicting the political alignment of Twitter users, ultimately predicting political affiliations at a 91% accuracy rate.

Skoric *et al*. (2012) suggest that the context in which the elections take place also is important. In particular, they find that concerns such as media freedoms, competitiveness of the elections and specifics of the electoral system may lead to certain over- and underestimations of voting sentiment when using Twitter tweets to predict votes.

However, there has been some controversy over the issue that the number of Twitter tweets can be used to predict elections (e.g. Tumasjan *et al*., 2010; Jungherr *et al*., 2012), in particular German elections. Apparently, counting occurrences of party names in Twitter messages is not sufficient. Chung and Mustafaraj (2011) and Gayo-Avello *et al*. (2011) found that simple counting and other approaches do not work well in predicting US Senate elections. However, recently, Sang and Bos (2012) in an analysis of a Dutch election, illustrate that substantial 'tuning' of Twitter messages can be used to improve prediction. As a result, researchers have suggested including both successes and failures in using Twitter as a tool for election prediction (e.g. Gayo-Avello, 2012).

Likely because of the variable results, models that include Twitter variables and other variables have been developed. For example, Tsakalidis *et al*. (2015) generated models based on Twitter and poll information to predict the outcome of elections in the EU (Germany, Netherlands and Greece). Their approach included capturing the number of times that different parties are mentioned in the Twitter tweets about the elections in particular countries. In addition, they used sentiment analysis to assign a sentiment value to each tweet. They find that the use of the Twitter variables generates statistically significantly better models than just using poll information.

8.2. Predicting the Spread of Disease

Sadilek *et al*. (2012) developed an approach to finding Twitter messages that are health-related tweets. Ritterman *et al*. (2009) used Twitter in conjunction with a prediction market in order to develop a prediction of a swine flu pandemic. Culota (2010) used keywords derived from Twitter messages to identify 'influenza'-related messages in order to determine the diffusion of the flu. Similarly, Chew and Eysenbach (2010) found that the use of the term 'H1N1' increased from 8.8% to 40.5% during the 2009 flu pandemic.

More recently, Li and Cardie (2013) used Twitter data as the basis of a 'real-time flu reporting system' that could be used to predict flu epidemics. Using a spatio-temporal unsupervised Bayesian algorithm, the system allows prediction of a flu breakout.

8.3. Predicting the Stock Market

Bollen *et al*. (2010) investigated the use of Twitter data to gather the 'mood' and use that information to facilitate stock market prediction. A similar approach also was used by Mittal and Goel (2012), who obtained similar but not as effective results.

8.4. Predicting Books, Movies, Music, iPhones and other Sales

Asur and Huberman (2010) found that using sentiment information after a movie was released allowed them to improve their prediction of movie revenues. Bhave *et al*. (2015) also found that Twitter sentiment analysis can contribute to the accuracy of predictions of movie success.

DOI: 10.1002/isaf

Although they did not analyse Twitter, Dhar and Chang (2009) analysed blogs and found that the volume of blog posts about a particular artist or album was positively correlated with future album music sales. Vossen (2013) specifically investigated the use of Twitter and found that Twitter message data was highly correlated with music sales. Vossen used messages that included either the artist's name or the album title in the tweet 2 weeks prior to or 1 week after the particular album's release date.

Using blog data, Gruhl *et al*. (2005) found that chatter information could be used to predict book sales. In a related study, Lassen *et al*. (2014) was able to use iPhone tweets to predict iPhone sales.

Although Twitter tweet information appears to be directly correlated to sales of books, movies, iPhones and music, Dijkman *et al*. (2015) find that the same does not hold for items that get less attention on social media. In order to more readily analyse other products, Dijkman *et al*. (2015) broke the tweets into additional subcategories to facilitate their analysis: job advertisement; product advertisement; positive or negative customer experience report; response to a customer experience report; daily chatter; factual statements about something that was bought; requesting information about the company or its products; pointing out or providing information or advice about the company or its products; news broadcast about the company; and other. Ultimately, such classifications are likely to depend on the particular use of the tweets and the products of interest.

8.5. Predicting Citations

Eysenbach (2011) used Tweets about the *Journal of Medical Internet Research* to examine citations, finding that highly tweeted articles were 11 times more likely to be cited than less tweeted articles. Unfortunately, most academic journals have limited number of Twitter tweets about them and even fewer mention specific articles. In addition, many academic journals are named after generic subject areas (e.g. 'decision support systems'), as a result a search for the journal turns up a large number of instances of the generic instance.

8.6. Predicting Soccer (Football) Matches

Bothos *et al*. (2010) and others have analysed the content of social media (e.g. blogs, Twitter and others) to predict the outcome of different events, including the 2010 World Cup. As a more recent example, Radosavljevic *et al*. (2014) analysed Tumblr blog posts in order to predict the relative strength of each country in the World Cup. After narrowing the blog posts based on a set of hashtags, they analysed the occurrence of both team mentions and player mentions. Using that data they developed a model called 'Goalr' that they used to predict the World Cup outcomes, including the outcomes of the matches in group play.

8.7. Predicting Accounting Estimates

As part of developing accounting financial statements, 'accounting estimates' are developed. As an example, companies need to generate estimates of expenses for product warranties, residual values and goodwill. One approach to gathering information for such estimates is to use Twitter. For example, information can be sought regarding a particular company and a particular product. Positive or negative information can be ascertained using a range of searches such as 'Xsucks'. Using this approach, qualitative information generated from Twitter can be captured and used to predict the amounts of expenses a company may face in the specific category.

DOI: 10.1002/isaf

9. CAUSATION AND TWITTER INFORMATION DISCLOSURES

The term ‘Twitter mining’ suggests that Twitter tweets do not affect the events being analysed – information is just analysed, after the fact. This raises the question of whether Twitter information influences or ‘causes’ the events or whether the resulting tweets are simply being mined as information. This section examines some of the issues that suggest that information in Twitter tweets, retweets, and so on is actually causing events, rather than simply providing information about the events.

9.1. Stock Market Events

The efficient markets hypothesis (e.g. Fama, 1970) suggests that all available information about a stock is fully reflected in the price. Historically, conventional information sources have been critical to gathering and diffusing that information. However, social media, such as Twitter, disintermediates many conventional information sources, bringing information sources in direct contact with others whose actions ultimately influence stock prices: newsmakers do not go through news services; they themselves break news on Twitter. Thus, researchers, such as Yu *et al*. (2013) found that social media has a stronger effect on market returns and risk than other news media. In addition, they found a strong interaction effect between social media and conventional media.

As a result, it is not surprising that some commentators (e.g. Viswanathan, 2013) have suggested that social media disclosures impact stock price. There are at least five ways that Twitter can provide new information to the markets:

- **News makers provide ‘new’ information**. Since Twitter can gather information directly from news producers, Twitter can provide the market new information. For example, Viswanathan (2013) indicated when Icahn tweeted that he had changed his position in Apple, the stock market responded, changing the price of Apple. In this case, the announcement came directly from Icahn and was information for the markets.
- **Firm-related information**. Stock market prices capture information about firms. Some aspects of that information can show up directly in Twitter and other social media as customers complain or praise a product. As another example, supply chain information can appear on Twitter, potentially influencing firm value. Since this information can occur on Twitter and other social media before it occurs in other conventional news sources, this twitter information about particular firms can be captured and embedded in stock prices. As an example, Lee *et al*. (2015) investigated product information that can be captured directly by social media and ultimately affect the market price.
- **Twitter information provides increased ‘investor recognition**’. Recently, Prokofieva (2014) used Merton’s (1987) ‘investor recognition hypothesis’ to frame one view of the effects of Twitter on stock markets. In particular, Merton (1987) suggested that firm value is increasing in the extent of investor recognition of the firm (holding fundamentals constant). As a result, Twitter and other social media can provide one approach to increase investor recognition as users tout particular information about companies and increase that recognition. Typically, it is assumed for the investor recognition hypothesis that the information being touted (stock or characteristics of the stock) needs to have low visibility: touting is likely to involve real information that the touter thinks is not visible enough or has not received enough attention.

  Touting for investor recognition, may be done by the particular companies or other market participants, such as shareholders. In either case there can be multiple motivations for the touting. However, it is likely that the touting is aimed at increasing the value of the stock being touted.

DOI: 10.1002/isaf

- **Twitter information from a hijacked account**. Lee (2013) reported on an instance where an Associated Press Twitter account was hijacked. False information about an injury to President Obama and an attack on the White House led to a 1% drop in the stock markets shortly after the announcement. Although the stock market rapidly rebounded when the information was determined to be false, this demonstrated the potential direct impact. Although the information was false, it came from a credible source that suggested that the attack was legitimate information.
- **Social media provides spam information**. Not all tweets are from influential market sources or even from legitimate sources. Frieder and Zittrain (2008) found that 'spam works'. In particular, they found that stock prices have been manipulated using spam emails; that is, stocks experience a significantly positive return on days prior to heavy touting via spam. Further, they also found that the volume of trading responds positively and significantly to 'heavy touting' and spam events (Hulbert, 2015).

  If stock prices can be manipulated by spam emails then it is likely that Twitter tweet or retweet information can do the same thing. A Twitter spammer would tweet either positive or negative information about the particular firm, depending on their strategy. Since people may treat the information as trading information, it is likely that, prior to heavy touting on Twitter, stocks can experience significant positive return and it is likely that volume of trading for some stocks is related to heavy touting. Spammers would not need to be robots. Even individuals might send tweets pushing particular aspects of some stock, hoping to increase its value.
- **Summary**. Touting easily integrates into Twitter and other social media for a number of applications. In stock market activity, touting is a part of both investor recognition and spam strategies. Although there are likely to be some differences emerging over time, touting is similar in both settings, with touting typically aimed at affecting the stock price. In each setting, touting could include touting legitimate information, although the importance in the messages of that legitimate information may differ. With spam, touting may not involve legitimate information or the information may be false. Further, spam tweets may only involve 'expectations' of the stock price.

9.2. Elections, Products, and So On

Similar to stock markets, election information about candidates or product information potentially could be pumped up or deflated down. Further, information that has had limited distribution, information from hijacked accounts and spam information could be investigated. As a result, information from Twitter can be used to cause a change in elections, a change in sales, and so on in each of the five settings discussed in Section 9.1.

Similarly, although our discussion of touting is aimed directly at stock market applications, touting may be done in a variety of applications. For example, a tweet might tout an electoral candidate or a tweet might tout a particular product that is being sold. Tweets could tout a potential winner of a soccer match, and other applications.

10. SUMMARY, CONTRIBUTIONS AND EXTENSIONS

This paper has investigated the notion of 'Twitter mining'. I have analysed issues such as what can be mined, how to analyse semantic message content and what kinds of additional information Twitter contains. This paper also summarized some applications used for prediction, discovery and the ability of Twitter information to cause particular events.

10.1. Contributions

This paper has summarized some of the literature of Twitter mining, providing a number of references to the Twitter mining literature that can be used to facilitate research into Twitter mining. This paper also analysed a number of methodology issues related to Twitter mining; for example, issues in sentiment analysis. In addition, this paper investigated a number of different types of applications, some of which could be extended to other settings. Finally, this paper expanded on the notions of touting.

10.2. Extensions

This paper can be extended in a number of different directions. First, data from additional sources, such as Facebook, could be integrated into the mining of Twitter data. For example, O'Leary (2012) investigated some of the relationships between the two data sources for privacy legislation. Second, other applications that use Twitter data that are not included here because of scope could be analysed. Although this paper references roughly 90 different sources, the literature is growing rapidly. Third, additional applications of Twitter data are likely to emerge over time. Future research could analyse those applications. Fourth, rather than limiting the analysis to Twitter, future research could analyse Tumblr, Stocktwits and other similar sources. Finally, future research could examine some of the effects on privacy of mining Twitter tweets.

ACKNOWLEDGEMENTS

An earlier version of this paper was presented at DSS 2.0 in Paris, June 2014. I would like to acknowledge the comments of the anonymous referees. In addition, I would like to thank the participants for their comments on the paper at the presentation.

REFERENCES

Acar A, Muraki Y. 2011. Twitter crisis for communication lessons learned from Japan's tsunami disaster. *Web-based Communities* **7**(3): 392–402.

Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*. Association for Computational Linguistics: Stroudsburg, PA; 30–38.

Allan J, Papka R, Lavrenko V. 1998. On-line new event detection and tracking. In *SIGIR '98, Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM: New York, NY; 37–45.

Asur S, Huberman BA. 2010. Predicting the future with social media. http://arxiv.org/PS_cache/arxiv/pdf/1003/1003.5699v1.pdf (accessed 25 August 2015).

Baccianella S, Esuli A, Sebastiani F. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *In LREC* **10**: 2200–2204.

Bakshy E, Mason W, Hofman J, Watts D. 2011. Everyone's an influencer: quantifying influence on Twitter. In *WSDM '11, Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. ACM: New York, NY; 65–74.

Bhave A, Kulkarni H, Biramane V, Kosamkar P. 2015. Role of different factors in predicting movie success. In *2015 International Conference on Pervasive Computing (ICPC)*. IEEE: Piscataway, NJ; 1–4.

Bollen J, Mao H, Zeng X. 2010. Twitter mood predicts the stock market. http://arxiv.org/pdf/1010.3003.pdf?iframe=true&width=90%25&height=90%25 (accessed 25 August 2015).

DOI: 10.1002/isaf

Bothos E, Apostolou D, Mentzas G. 2010. Using social media to predict future events with agent-based markets. *IEEE Intelligent Systems* **25**(6): 50–58.

Cha M, Haddadi H, Benevenuto F, Gummadi KP. 2010. Measuring user influence in Twitter: the million follower fallacy. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence: Palo Alto, CA; 10–17.

Chew C, Eysenbach G. 2010. Pandemics in the age of Twitter: content analysis of tweets during the H1N1 outbreak. *PLoS ONE* **5**(11): e14118.

Chung J, Mustafaraj E. 2011. Can collective sentiment expressed on Twitter predict political elections?. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence: Palo Alto, CA; 1770–1771.

Conover M, Goncalves B, Ratkiewicz J, Flammini A, Menczer F. 2012. Predicting the political alignment of Twitter users. http://cnets.indiana.edu/wp-content/uploads/conover_prediction_socialcom_pdfexpress_ok_version.pdf (accessed 25 August 2015).

Culota A. 2010. Toward detecting influenza epidemics by analyzing Twitter messages. In *SOMA '10, Proceedings of the First Workshop on Social Media Analytics*. ACM: New York, NY; 115–122.

DeJong G. 1979. Prediction and substantiation: a new approach to natural language. *Cognitive Science* **3**: 251–273.

Dhar V, Chang E. 2009. Does chatter matter? The impact of user generated content on music sales. *Journal of Interactive Marketing* **23**(4): 300–307.

DiGrazia J, McKelvey K, Bollen J, Rojas F. 2013. More tweets, more votes: Social media as a quantitative indicator of political behavior. *PLoS ONE* **8**(11): e79449.

Dijkman R, Ipeirotis P, Aertsen F, van Helden R. 2015. Using Twitter to predict sales: a case study. http://arxiv.org/ftp/arxiv/papers/1503/1503.04599.pdf (accessed 25 August 2015).

Ediger D, Jiang K, Riedy J, Bader D, Corley C, Farber R, Reynolds W. 2010. Massive social network analysis: mining Twitter for the social good. In *ICPP '10 Proceedings of the 2010 39th International Conference on Parallel Processing*. IEEE Computer Society: Washington, DC; 583–593.

Eysenbach G. 2011. "Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research* **13**(4): e123.

Fama EF. 1970. Efficient capital markets: a review of theory and empirical work. *Journal of Finance* **25**(2): 383–417.

Fellbaum C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press: Cambridge Massachusetts.

Fox M. 2014. Yelp helps pinpoint food poisoning at NYC restaurants. http://www.nbcnews.com/health/health-news/yelp-helps-pinpoint-food-poisoning-nyc-restaurants-n112266 (accessed 25 August 2015).

Frieder L, Zittrain J. 2008. Spam works: evidence from stock touts and corresponding market activity," Berkman Center Research Publication No. 2006-11(2007), 13 Hastings Communications and Entertainment Law Journal 479.

García-Cumbreas MA, García-Vega M, Martínez-Santiago F, Peréa-Ortega JM. 2010. SINAI at WEPS-3: online reputation management. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.174.3058&rep=rep1&type=pdf (accessed 26 August 2015).

Gayo-Avello D. 2012. A balanced survey on election prediction using Twitter data. http://arxiv.org/pdf/1204.6441.pdf (accessed 26 August 2015).

Gayo-Avello D, Metaxas P, Mustafaraj E. 2011. Limits of electoral predictions using Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence: Palo Alto, CA; 490–493.

Go A, Huang L, Bhayani R. 2009. Sentiment analysis of Twitter data. http://www-nlp.stanford.edu/courses/cs224n/2009/fp/3.pdf (accessed 26 August 2015).

Gruhl D, Guha R, Kumar R, Novak J, Tomkins A. 2005. The predictive power of online chatter. In *KDD '05, Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. ACM: New York, NY; 78–87.

Hayes PJ, Weinstein S. 1990. Construe-TIS: a system for content-based indexing of a database of news stories. In *IAAI '90 Proceedings of the Second Conference on Innovative Applications of Artificial Intelligence*. AAAI Press: Palo Alto, CA; 49–64.

Hu M, Liu B. 2004. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004, full paper)*, Seattle, Washington, USA.

Huang C. 2011. Facebook and Twitter key to Arab spring uprisings: report. *The National*, 6 June. http://www.thenational.ae/news/uae-news/facebook-and-twitter-key-to-arab-spring-uprisings-report (accessed 26 August 2015).

Huberman BA, Romero DM, Wu F. 2008. Social networks that matter: Twitter under the microscope. http://arxiv.org/pdf/0812.1045.pdf (accessed 25 August 2015).

Hughes AL, Palen L. 2009. Twitter adoption and use in mass convergence and emergency events. In *Proceedings of the 6th International ISCRAM Conference – Gothenburg, Sweden*, Landgren J, Jul S (eds). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.156.8385&rep=rep1&type=pdf (accessed 26 August 2015).

Hulbert M. 2015. Why the Twitter hoax suggests that the stock market is near a top. http://www.marketwatch.com/story/what-the-twitter-hoax-says-about-the-stock-market-2015-07-14 (accessed 26 August 2015).

Jansen BJ, Zhang M, Sobel K, Chowdury A. 2009. Twitter power: tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology* **60**(11): 2169–2188.

Java A, Finin T, Song X, Tseng B. 2007. Why we Twitter: understanding microblogging usage and communities. http://aisl.umbc.edu/resources/369.pdf (accessed 26 August 2015).

Jungherr A, Jürgens P, Schoen H. 2012. Why the party won the German election of 2009 or the trouble with predictions: a response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. “Predicting elections With Twitter: what 140 characters reveal about political sentiment”. *Social Science Computer Review* **30**: 229–234.

Jurgens D. 2013. That’s what friends are for: inferring location in online social media platforms based on social relationships. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. AAAI Press: Palo Alto, CA; 273–282.

Kwak H, Lee C, Park H, Moon S. 2010. What is Twitter, a social network or a news media? In *WWW ’10 Proceedings of the 19th International Conference on World Wide Web*. ACM: New York, NY; 591–600.

Lassen NB, Madsen R, Vatrapu R. 2014. Predicting iPhone sales from iPhone tweets. In *2014 IEEE 18th International Enterprise Distributed Object Computing Conference (EDOC)*. IEEE: Piscataway, NJ; 81–90.

Lee E. 2013. AP Twitter account hacked in market-moving attack. http://www.bloomberg.com/news/articles/2013-04-23/dow-jones-drops-recovers-after-false-report-on-ap-twitter-page (accessed 26 August 2015).

Lee L, Hutton A, Shu S. 2015. The role of social media in the capital market: evidence from consumer product recalls. *Journal of Accounting Research* **53**: 367–404.

Li J, Cardie C. 2013. Early stage influenza detection from Twitter. http://arxiv.org/abs/1309.7340 (accessed 26 August 2015).

Liebowitz M. 1980. Generalization and memory in an integrated understanding system. PhD dissertation, Yale University.

Lohr S. 2010. Library of Congress will save tweets. *New York Times*, 14 April. http://www.nytimes.com/2010/04/15/technology/15twitter.html (accessed 26 August 2015).

Loughran T, McDonald B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance* **66**(1): 35–65.

Mandel B, Culotta A, Boulahanis J, Stark D, Lewis B, Rodrigue J. 2012. A demographic analysis of online sentiment during hurricane Irene. In *LSM ’12 Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics: Stroudsburg, PA; 27–36.

Maron D. 2014. Tweets identify food poisoning outbreaks. *Scientific American*, 20 August. http://www.scientificamerican.com/podcast/episode/tweets-identify-food-poisoning-outbreaks/ (accessed 26 August 2015).

Mendoza M, Poblete B, Castillo C. 2010. Twitter under crisis: can we trust what we RT?. In *SOMA ’10, 1st Workshop on Social Media Analytics*. ACM: New York, NY; 71–79.

Merton R. 1987. A simple model of capital market equilibrium with incomplete information. *The Journal of Finance* **42**(3): 483–510.

Mittal A, Goel A. 2012. Stock prediction using Twitter sentiment analysis. http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf (accessed 26 August 2015).

Mohammad S, Dunne C, Dorr B. 2009. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2*. Association for Computational Linguistics: Stroudsburg, PA; 599–608.

Mueller ET. 2002. *Story understanding*. In *Encyclopedia of Cognitive Science*. Macmillan: London.

Mueller ET. 2004. Understanding Script-based stories using commonsense reasoning. *Cognitive Systems Research* **5**(4): 307–340.

Nguyen T-M, Koshikawa K, Kawamura T, Tahara Y, Ohsuga A. 2011. Building earthquake semantic network by mining human activity from Twitter. In 2011 IEEE International Conference on Granular Computing. IEEE: Piscataway, NJ; 496–501.

DOI: 10.1002/isaf

Nishida K, Banno R, Fujimura K, Hoshide T. 2011. Tweet classification by data compression. In *DETECT '11, Proceedings of the 2011 International Workshop on Detecting and Exploiting Cultural Diversity on the Social Web*. ACM: New York, NY; 29–34.

O'Leary DE. 2011. Blog mining-review and extensions: 'From each according to his opinion'. *Decision Support Systems* **51**(4): 821–830.

O'Leary DE. 2012. Computer-based political action: the battle and Internet blackout over PIPA. *IEEE Computer* **45**(7): 64–72.

O'Leary DE. 2013a. 'Big data', the 'Internet of Things' and the 'Internet of Signs. *Intelligent Systems in Accounting, Finance and Management* **20**(1): 53–65.

O'Leary DE. 2013b. Knowledge discovery for continuous financial assurance using multiple types of digital information. In *Contemporary Perspectives in Data Mining, Volume 1*, Lawrence KD, Klimberg R (eds). Information Age Publishing: Charlotte, NC; 103–122.

Papka R. 1999. On-line new event detection, clustering and tracking. PhD dissertation, University of Massachusetts–Amherst.

Prokofieva M. 2014. Twitter-based dissemination of corporate disclosure and the intervening effects of firms' visibility: evidence from Australian listed companies. *Journal of Information Systems* **29**(2): 107–136.

Radosavljevic V, Grbovic M, Djuric N, Bhamidipati N. 2014. Large-scale World Cup 2014 outcome prediction based on Tumblr posts. In *KDD Workshop on Large-Scale Sports Analytics*, Sydney, Australia. http://www.large-scale-sports-analytics.org/Large-Scale-Sports-Analytics/LastYearSubmissions_files/paperID10.pdf.

Ritterman J, Osborne J, Klein E. 2009. Using prediction markets and Twitter to predict a swine flu pandemic. In *1st International Workshop on Mining Social Media*, Seville, Spain. http://homepages.inf.ed.ac.uk/miles/papers/swine09.pdf (accessed 26 August 2015).

Russell M. 2011. *21 Recipes for Mining Twitter*. O'Reilly Media: Sebastopol, CA.

Russell MA. 2013. *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. O'Reilly Media: Sebastopol, CA.

Sadilek A, Kautz HA, Silenzio V. 2012. Modeling spread of disease from social interactions. In *Sixth AAAI International Conference on Weblogs and Social Media*. AAAI Press: Palo Alto, CA; 322–329.

Sang ETK, Bos J. 2012. Predicting the 2011 Dutch senate election results with Twitter. In *13th Conference of the European Chapter of the Association for Computational Linguistics: Proceedings of the Workshop on Semantic Analysis in Social Media*. Association for Computational Linguistics: Stroudsburg, PA; 53–60.

Shen W, Wang J, Luo P, Wang M. 2013. Linking named entities in tweets with knowledge base via user interest modeling. In *KDD'13 Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM: New York, NY; 68–76.

Suh B, Hong L, Pirolli P, Chi E. 2010. Want to be retweeted? Large scale analytics on factors impacting retweet in a Twitter network. In *2010 IEEE Second International Conference on Social Computing (SocialCom)*. IEEE: Piscataway, NJ; 177–184.

Skoric M, Poor N, Achananuparp P, Lim EP, Jiang J. 2012. Tweets and votes: a study of the 2011 Singapore general election. In *2012 45th Hawaii International Conference on System Science (HICSS)*. IEEE: Piscataway, NJ; 2583–2591.

Stone P, Dunphy D, Smith M, Ogilvie D. 1966. *The General Inquirer: A Computer Approach to Content Analysis*, Vol. **08**. MIT Press: Cambridge.

Sutter J. 2009. Swine flu creates controversy on Twitter," 30 April. http://edition.cnn.com/2009/TECH/04/27/swine.flu.twitter/index.html (accessed 27 August 2015).

Swartz J. 2009. Social media like Twitter change customer service. *USA Today*, 18 November. http://usatoday30.usatoday.com/tech/news/2009-11-18-twitterserve18_ST_N.htm (accessed 27 August 2015).

Teufl P, Kraxberger S. 2011. Extracting semantic knowledge from Twitter. In *Electronic Participation*, Tambouris E, Macintosh A, de Bruijn H (eds). Lecture Notes in Computer Science, vol. **6847**. Springer: Berlin; 48–59.

Thelwall M, Buckley K, Paltoglou G. 2011. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology* **62**(2): 406–411.

Tsakalidis A, Papadopoulos S, Cristea A, Kompatsiaris Y. 2015. Predicting elections for multiple countries using Twitter and polls. *IEEE Intelligent Systems* **30**(2): 10–17.

Tumasjan A, Sprenger T, Sandner P, Welpe I. 2010. Predicting elections with Twitter: what 140 characters reveal about political sentiment. In *Proceedings of the Fourth AAAI Conference on Weblogs and Social Media*. AAAI Press: Palo Alto, CA; 178–185.

Vargas-Vera M, Celjuska D. 2004. Event recognition on news stories and semi-automatic population of an ontology. In *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE Computer Society; 615–618.

Vieweg S, Hughes A, Starbird K, Palen L. 2010. Microblogging during two natural hazard events: what Twitter might contribute to situation awareness. In *CHI '10, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM: New York, NY; 1079–1088.

Viswanathan B. 2013. Does social media affect capital markets? 10 September. http://www.forbes.com/sites/quora/2013/09/10/does-social-media-affect-capital-markets/ (accessed 27 August 2015).

Vossen R. 2013. Does chatter matter? Predicting music sales with social media. http://www.basicthinking.de/blog/wp-content/uploads/2013/06/Does-Chatter-Matter.pdf (accessed 24 August 2015).

Weng J, Lim E-P, Jiang J. 2010. TwitterRank: finding topic-sensitive influential twitters. In *WSDM '10, Proceedings of the Third ACM International Conference on Web Search and Data Mining*. ACM: New York, NY; 261–270.

Westermann U, Jain R. 2007. Toward a common event model for multimedia applications. *IEEE Multimedia* **14**(1): 19–29.

Wexler A. 2013. Twitter triumph: 100% of Fortune 500 soon will tweet. 7 November. http://www.huffingtonpost.com/adam-wexler/twitter-triumph-100-of-fo_b_4235213.html (accessed 27 August 2015).

Wilson T, Wiebe J, Hoffmann P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT '05, Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics: Stroudsburg, PA; 347–354.

Yalamanchi D. 2011. Sideffective – system to mine patient reviews: sentiment analysis. MS thesis, Rutgers University, New Brunswick, NJ.

Yu Y, Duan W, Cao Q. 2013. The impact of social and conventional media on firm equity value: a sentiment analysis approach. *Decision Support Systems* **55**: 919–926.

Zhang D, Liu Y, Lawrence R, Chenthamarakshan V. 2011. Transfer latent semantic learning: microblog mining with less supervision. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*. AAAI Press: Palo Alto, CA; 561–566.

DOI: 10.1002/isaf